Applied Concept Paper #3: Classifying Clinical Data

Introduction:

The health concerns and mortality related to COVID-19 among patients varies based on the severity of disease. During hospitalization the vitals and laboratory measurements are taken and these results can be useful for predicting severity of disease leading to efficient and timely resource allocation preventing adverse health outcomes. [1] Many studies have proposed model to predict mortality and ventilator management of mild to moderate COVID-19. [2] Some studies developed model for the prediction of incubator and mechanical ventilation for patients in the emergency room. [3] Vitals and laboratory tests results are associated with severity of COVID-19 as has been shown in the applied concept paper1. This study aims to predict the severity of a patient based on the relevant predictors among the 22 vital and laboratory test results collected during 14 days since the hospital visit. An average value over 14 days of hospital stay for 22 labs and vitals were considered for model selection and performance evaluation.

Method:

This study applies two models in sequence, first model (Model1) identifies if the patient has mild or not-mild COVID-19. If the patient has not-mild COVID-19 then the second model (Model2) finds if he/she has moderate or severe COVID-19. The Model1 and Model2 were determined by evaluating performances of Logistic Regression, Decision Tree and Random Forest on the relevant training sets.

A pipeline was created comprising of a sequence of steps like preprocessing, feature selection and the estimator to automate the training, validation and testing process. This also ensures no information leak from the test and the validation set in any of the steps in the pipeline during the training process. The preprocessing step comprised of an imputation step using SimpleImputer that replaced the missing values with the means of respective features, standardization of features using StandardScaler, and feature selection with RFECV of the scikit-learn [4] library in python. Scaling was skipped for the Decision Tree and the Random Forest estimators during preprocessing as normalization is not required for a tree-based algorithm since it depends on partitioning of data to make predictions. The feature selection was done using recursive feature elimination with cross-validation. The hyperparameters of an estimator was tuned by exhaustive search using GridSearchCV function and cross_val_score function was invoked to obtain a better estimate of validation score during model selection.

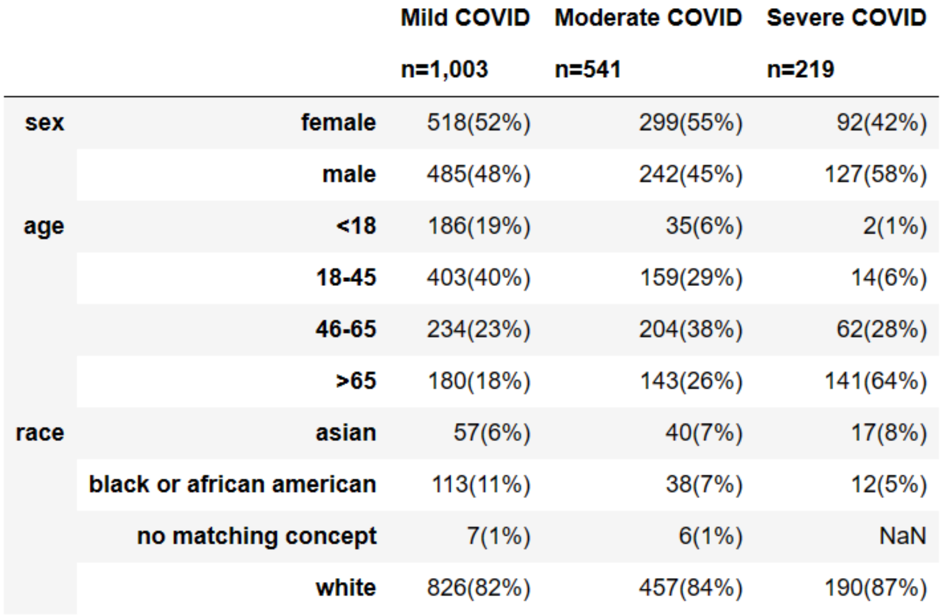
The data for this study comprised of patient id, demographic information like gender, race/ethnicity, age at the time of hospitalization, and the measurements for 22 vitals and laboratory tests taken within 14 days since the start of visit. The records of the patients were condensed by averaging readings for each measurement type by the day from the hospital visit, and subsequently by averaging over all of the 14 days. As a result, the dataset comprised of patient-id, demographic information, and the average readings of 22 labs/vitals over 14 days in the hospital.

The dataset for Model1 that predicts mild vs non-mild case was derived by relabeling the moderate and severe cases as non-mild, and that for Model2 was derived by filtering only the patients with moderate and severe cases. The dimension of dataset for Model1 was 1763 x 23 and for Model2 was 760 x 23, where 22 columns are features and one is the severity category, target variable. The data was split 80/20 into train and test sets in stratified fashion based on the severity category. Nested and stratified cross-validation was performed during hyperparameter tuning with 5 folds (inner loop), and model selection with 2 folds (outer loop). The average cross-validation accuracy provides a good estimate of the score if the model is tuned for hyperparameters and used on unseen data, therefore both Model1 and Model2 were selected based on the nested cross-validation approach. The scoring metrics computed to estimate model’s e performance were accuracy, f1 score and roc_auc.

Result:

The demographic information by the disease severity (mild, moderate and severe) is summarized in Table 1. The dataset comprises of 1763 patients with 1,003 mild cases and 760 moderate and severe cases. The demographic distribution indicates that a higher proportion of mild and moderate patients are female, but that of severe are male. The severe patients comprise mostly (64%) of older population with age >65. Also, the data represents mostly whites with approx. 80% of the records in all of the severity categories.

Table 1 Summarized demographic (sex, age, and race) information of 1763 patients in the sample categorized as Mild COVID, Moderate COVID and Severe COVID. No matching concept implies unknown race.



		Mild COVID	Moderate COVID	Severe COVID
		n=1,003	n=541	n=219
sex	female	518(52%)	299(55%)	92(42%)
	male	485(48%)	242(45%)	127(58%)
age	<18	186(19%)	35(6%)	2(1%)
	18-45	403(40%)	159(29%)	14(6%)
	46-65	234(23%)	204(38%)	62(28%)
	>65	180(18%)	143(26%)	141(64%)
race	asian	57(6%)	40(7%)	17(8%)
	black or african american	113(11%)	38(7%)	12(5%)
	no matching concept	7(1%)	6(1%)	NaN
	white	826(82%)	457(84%)	190(87%)

As mentioned in the method, 22 laboratory and vitals were averaged over 14 days of hospital visit to obtain the dataset with 22 features and the COVID-19 severity categories for the 1763 patients. The laboratory and vital names were abbreviated for convenience and is provided in Table 2.

Table 2 Abbreviated column names

Feature names: Abbreviated names
'Body temperature': 'Body temp', 'Diastolic blood pressure': 'DBP', 'Heart rate': 'HR', 'Body weight': 'BW', 'Systolic blood pressure': 'SBP', 'Oxygen saturation in Arterial blood': 'SaO2', 'Respiratory rate': 'RR',

```
 'Erythrocytes [#/volume] in Blood by Automated count': 'Erythrocytes',
 'Hemoglobin [Mass/volume] in Blood': 'Hemoglobin',
 'Hematocrit [Volume Fraction] of Blood by Automated count': 'Hematocrit',
 'MCHC [Mass/volume] by Automated count': 'MCHC',
 'Glomerular filtration rate/1.73 sq M.predicted [Volume Rate/Area] in Serum, Plasma or Blood by
Creatinine-based formula (MDRD)': 'MDRD GFR',
 'Protein [Mass/volume] in Serum or Plasma': 'Protein',
 'Alkaline phosphatase [Enzymatic activity/volume] in Serum or Plasma': 'Alkaline phosphatase',
 'Alanine aminotransferase [Enzymatic activity/volume] in Serum or Plasma': 'Alanine aminotransferase',
 'Aspartate aminotransferase [Enzymatic activity/volume] in Serum or Plasma': 'Aspartate aminotransferase',
 'Leukocytes [#/volume] in Blood by Automated count': 'Leukocytes',
 'MCV [Entitic volume] by Automated count': 'MCV',
 'MCH [Entitic mass] by Automated count': 'MCH',
 'Platelets [#/volume] in Blood by Automated count': 'Platelets',
 'Albumin [Mass/volume] in Serum or Plasma': 'Albumin',
 'Bilirubin.total [Mass/volume] in Serum or Plasma': 'Bilirubin.total'
```

The aim of this study is to predict the severity of COVID-19 for a patient. This is done in a two-step process: first the Model1 predicts if the patient has mild or non-mild COVID-19. In case the patient has non-mild COVID-19 then Model2 will predict if its moderate or severe. The details on the dataset preparation for the two models is provided in the method. Three estimators for the binary classification were trained and evaluated through cross-validation for Model1 and Model2. The pipelines of the logistic regression, decision tree and random forest are shown in Table 3. As seen in the table, the datasets were preprocessed before training and validation. For all three estimators the missing values were replaced by the respective mean value of the feature. Scaling was implemented only for logistic regression. Next, the features that are relevant in classifying the severity were selected using RFECV, before training the estimator on the preprocessed dataset. The relevant features selected for the three estimators for the two models are provided in Table 4. It is interesting to see that decision tree found only protein among the 22 features to be relevant for classifying the mild vs non-mild patients. Significant reduction in dimension was achieved employing the feature selection method.

Table 3 Pipeline of steps

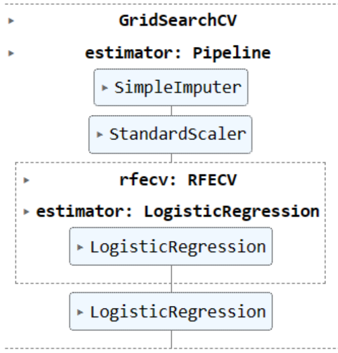
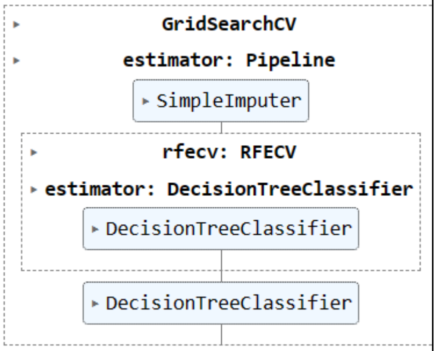
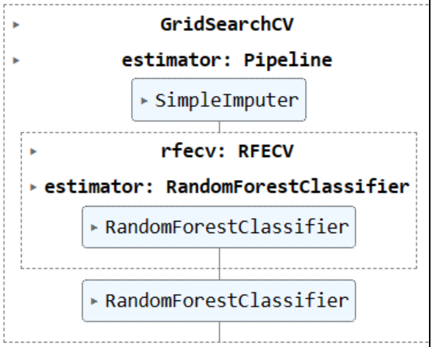
Logistic Regression	Decision Tree	Random Forest
 ▸ GridSearchCV ▸ estimator: Pipeline ▸ SimpleImputer ▸ StandardScaler ▸ rfecv: RFECV ▸ estimator: LogisticRegression ▸ LogisticRegression ▸ LogisticRegression	 ▸ GridSearchCV ▸ estimator: Pipeline ▸ SimpleImputer ▸ rfecv: RFECV ▸ estimator: DecisionTreeClassifier ▸ DecisionTreeClassifier ▸ DecisionTreeClassifier	 ▸ GridSearchCV ▸ estimator: Pipeline ▸ SimpleImputer ▸ rfecv: RFECV ▸ estimator: RandomForestClassifier ▸ RandomForestClassifier ▸ RandomForestClassifier

Table 4 Feature selection using Recursive Feature Elimination with 5-fold Cross Validation

Estimators	**Model1 (mild vs non-mild)**	**Model2 (moderate vs severe)**
Logistic Regression	'BW', 'MDRD GFR', 'Leukocytes', 'Platelets', 'Protein'	'Alanine aminotransferase', 'Albumin', 'Aspartate aminotransferase', 'Bilirubin.total', 'Body temp', 'BW', 'DBP', 'MDRD GFR', 'Hematocrit', 'MCH', 'SaO2', 'Platelets', 'Protein', 'RR', 'SBP'
Decision Tree	'Protein'	'Alanine aminotransferase', 'Body temp', 'DBP', 'SBP'
Random Forest	'Alanine aminotransferase', 'Albumin', 'Alkaline phosphatase', 'Aspartate aminotransferase', 'Bilirubin.total', 'BW', 'Erythrocytes', 'MDRD GFR', 'HR', 'Hematocrit', 'Hemoglobin', 'Leukocytes', 'MCH', 'MCHC', 'MCV', 'SaO2', 'Platelets', 'Protein', 'RR'	'Alanine aminotransferase', 'Alkaline phosphatase', 'Aspartate aminotransferase', 'Body temp', 'BW', 'DBP', 'HR', 'Hemoglobin', 'MCH', 'SaO2', 'Platelets', 'SBP'

For both models the hyperparameters were tuned for all the three estimators to achieve an efficient classification. An exhaustive search over the specified parameter values were performed to obtain the optimal set for the three estimators, see Table 5 and Table 6. The best estimator for the classification task was determined by comparing the accuracy, F1 and ROC AUC scores for both of the models, see Table 7.

Table 5 Hyperparameter tuning: the parameters and the value range tuned for better performance

Logistic Regression	Decision Tree	Random Forest
C: [.0001, .001, .01, .1, 1.0, 10.0, 100.0, 1000.0, 10000.0] solver: ['liblinear'] penalty: ['l1', 'l2']	max_depth: [1,2,3,4,5,6,7,None] criterion: ['gini', 'entropy']	n_estimators: [10, 100, 1000] criterion: ['gini', 'entropy']
C: [.0001, .001, .01, .1, 1.0, 10.0, 100.0, 1000.0, 10000.0] solver: ['liblinear'] penalty: ['l2']		

Table 6 Optimal set of hyperparameters using exhaustive search over specified parameter values for an estimator using GridSearchCV. Both models selected the same set of hyperparameters.

Logistic Regression	Decision Tree	Random Forest
C: 1.0 solver: lbfgs penalty: l2	criterion: gini max_depth: None	n_estimators: 100 criterion: gini

Table 7 Scores (value ± standard deviation) of logistic regression, decision tree and random forest for binary classifications: 1. Mild vs Non-mild 2. Moderate vs Severe.

Estimators	**Model1 (mild vs non-mild)**			**Model2 (moderate vs severe)**		
	Accuracy	F1	AUC	Accuracy	F1	AUC
Logistic Regression	0.928 ± 0.011	0.909 ± 0.015	0.922 ± 0.022	0.699 ± 0.015	0.271 ± 0.010	0.695 ± 0.004
Decision Tree	0.989 ± 0.003	0.987 ± 0.003	0.989 ± 0.006	0.681 ± 0.033	0.316 ± 0.122	0.583 ± 0.032
Random Forest	0.991 ± 0.001	0.990 ± 0.002	0.995 ± 0.001	0.714 ± 0.000	0.215 ± 0.043	0.696 ± 0.019

Discussion:

The dataset for Model1 that classifies mild vs non-mild is almost balanced as the percentages of mild and non-mild patients in the dataset are 56.9% and 43.1% respectively. Consequently, accuracy score can be considered for selecting the best estimator. However, it was found that all three scores: accuracy, F1 and AUC unanimously supported random forest as the most efficient estimator for predicting mild vs non-mild as shown in Table 7. On the contrary, the dataset for Model2 has some amount of imbalance with 71.1% moderate and 28.9% severe cases. In case of class imbalance F1 and AUC are considered for model selection. Decision tree performs best for Model2 based on F1, however random forest and logistic regression performs better if AUC is considered. Hence, both logistic regression and random forest could be used for classification of moderate or severe cases. The performance of Model1 with optimized hyperparameter on the test set, and that of the Model2 also with optimized hyperparameter on the test set are given in Table 8.

The Model1 performs well on the test set with 0.997 F1 and AUC scores. However, we see that Model2 is biased and performs poorly. This is also evident from the confusion matrices and ROC curves for Model1 and Model2 shown in Figure 1 and Figure 2 respectively. Model1 identifies the mild and non-mild accurately, whereas Model2 does a decent job in identifying the moderate patients but performs poorly in labelling the severe patients. 32 out of 44 of the severe patients were classified incorrectly as moderate by the Model2 classifier. There are two possible ways to resolve this: To remove bias in Model2 we could use more complex model like deep learning, or the data we are using to train the Model2 might not have the right information to differentiate between moderate and severe. We learnt from ACP1 paper that the lab and vital values between the moderate and severe patients vary after 7 days since the start of the hospital visit, so a better data for training for Model2 could be obtained by averaging over the last 7 days.

This study is useful since the classification of severity is important for resource allocation and management, and proper initiatives could be taken to deter unwanted health outcomes for moderate and severe patients.

Table 8 Model1 and Model2 performances on the test set. Both models used optimized hyperparameter.

Estimator	**Model1 (mild vs non-mild)**		**Model2 (moderate vs severe)**	
	F1	AUC	F1	AUC

Random Forest	0.997	0.997	0.387	0.609

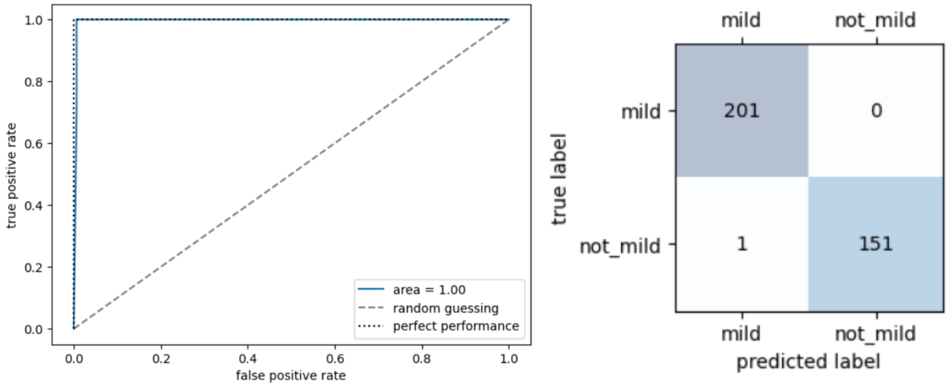


Figure 1 The receiver operating characteristic curve and the area under the curve (left), and the confusion matrix (right) for Model1 (mild vs non-mild), a random forest estimator with optimized parameters.

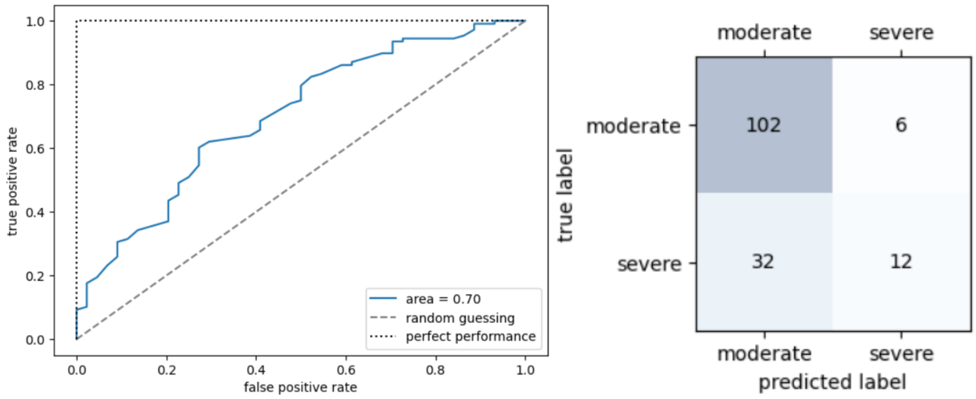


Figure 2 The receiver operating characteristic curve and the area under the curve (left), and the confusion matrix (right) for Model2 (moderate vs severe), a random forest estimator with optimized parameters.

Limitations:

The data predominantly represents white population which is 80% in all of the three (mild, moderate and severe) categories. As a result, the estimators are more suitable for severity prediction among whites and might perform poorly for patients of another race and ethnicity. The laboratory and vitals are sparse for mild patients, resulting in a lot of missing values for the mild patients. Imputation of the missing values with the average could result in less variation in feature values between the classes. This could impact the performance of the classification model. Also, the difference between the moderate and severe patients' vitals/labs results is more prominent towards the last 7 days of hospital visit. Consequently, using a data that averages the values over last 7 days could be a better dataset to train Model2 and could improve its F1 and AUC score.

References

[1] Ahmed Sameer Ikram and Somasundram Pillay, "Admission vital signs as predictors of COVID-19 mortality: a retrospective cross-sectional study," *BMC Emergency Medicine,* 2022.

[2] Rechtman, E., Curtin, P., Navarro, E., Nirenberg, S. & Horton, M. K, "Vital signs assessed in initial clinical encounters predict covid-19 mortality in a nyc hospital system.," *Sci. Rep.,* p. 21545.

[3] Kazuya Sakai, Kai Okoda, Mototsugu Nishii, Ryo Saji, Fumihiro Ogawa, Takeru Abe & Ichiro Takeuchi , "Combining blood glucose and SpO2/FiO2 ratio facilitates prediction of imminent ventilatory needs in emergency room COVID-19 patients," *Scientific Reports,* p. 22718, 2023.

[4] Pedregosa et al., "Scikit-learn: Machine Learning in Python," vol. 12, no. 85, pp. 2825-2830, 2011.